Efficient soil loss assessment for large basins using smart coded polygons

Efficient Soil Loss Assessment For Large Basins Using Smart Coded Polygons

Jinren Ni

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

Ao Wu

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

Tianhong Li

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

Yao Yue

Key Laboratory of Water and Sediment Sciences, Ministry of Education, Beijing, PR China; Department of Environmental Engineering, Peking University, PR China

Alistair GL Borthwick

School of Engineering, The University of Edinburgh, The King's Buildings, Edinburgh EH9 3JL, U.K.

Corresponding Author

Jinren Ni, Department of Environmental Engineering, Peking University, NO 5 Yiheyuan Road, Beijing 100871, PR China

Email: nijinren@iee.pku.edu.cn

Tianhong Li, Department of Environmental Engineering, Peking University, NO 5 Yiheyuan Road, Beijing 100871, PR China

23 Email: lth@pku.edu.cn

Abstract

Soil erosion is a severe ecological problem. Most conventional methodologies for soil-erosion assessment are appropriate for small or medium river basins. This paper presents an approach to soil-erosion intensity assessment in large basins, utilizing coded polygons identified by spatially overlapping gradation levels of primary environmental factors. Efficient assessment of soil-erosion intensity is achieved by matching the coded polygons to selected polygons pre-assigned to reference groups. A case study is presented for the soil-erosion assessment of the Yellow River Basin. It is found that the calculated and observed soil-erosion intensities are in close agreement for 86% of the total area. Sensitivity analysis indicates that acceptable results are obtained using a 5% sample of the original 9921 coded polygons, thus reducing substantially the computational load. Direct comparisons between the polygon codes in the reference and test groups show that uncertainty is reduced with respect to previous methods. This is confirmed by the reduction in information entropy from 7.49 to 1.33. The proposed approach should be of particular use in the cost-effective assessment of soil erosion in large basins.

Keywords

Coded polygons, Soil erosion assessment, Yellow River Basin, Information classification, Semi-quantitative model

1. Introduction

Soil erosion causes 84% of land degradation worldwide (Eswaran et al., 2001) and leads to other severe environmental problems such as river sedimentation and non-point pollution (Pimentel et al., 1995; UNEP, 2007; Telles et al., 2011). The global area of land degraded by water erosion covers nearly 1,100 Mha and is predominantly located in Asia and Africa (Oldeman, 1994). In China, the gross quantity of eroded soil exceeds 5 billion tons per year, accounting for about 8% of the World's total (Jing et al., 2005). The Second National Survey of Soil Erosion indicated that 37% of China's land area was affected by water and soil loss, with an even larger area undergoing soil erosion and deposition processes (Jing et al., 2005).

In the 20th Century, the primary factors influencing soil erosion were fully investigated, including precipitation, vegetation, soil type, and land management (Zingg, 1940; Smith and Whitt, 1948; Meyer, 1984). Several empirical models were proposed for assessing the status of soil erosion, based on knowledge of the environmental factors and physical processes involved. The Universal Soil Loss Equation (USLE) was proposed by the U.S. Department of Agriculture (Wischmeier and Smith, 1965; Meyer, 1984), and later revised as RUSLE (Renard et al., 1997). Although USLE/RUSLE has been used worldwide (Wang and Jiao, 1996; Biesemans et al., 2000; Li et al., 2010; Dabney et al., 2011; Xu et al., 2011), it is not always exactly applicable and has occasionally been misused (Wischmeier, 1976; Boardman, 2006). USLE works best

66 for regions in the USA (Stocking, 1995; Vrieling et al., 2002), with amendments
67 necessary for other areas. Moreover, the original USLE model was derived from plot
68 experiments and so is only directly applicable at plot-scale (Terranova et al, 2009;
69 Kinnell 2010). For large-scale applications, the study areas have to be separated into
70 cells or sub-basins until the resulting units are sufficiently small for USLE to be
71 correctly implemented (Millward and Mersey, 1999; Chen et al., 2011; Iroum et al,
72 2011; Shinde et al., 2011). Ideally, the parameters required for each unit should be
73 derived using 3S technology (Global Positioning System, Remote Sensing, and
74 Geographic Information System). Remote sensing can provide high-resolution images
75 and GIS enables rapid spatial analysis, incorporating the DEM dataset, slope
76 calculations, division of river basins, and so on. However, such data requirements are
77 presently beyond the capabilities of many developing countries in Asia and Africa
78 where soil erosion is particularly severe (Stocking, 1995; Ananda and Herath, 2003;
79 Vrieling, 2006). Physically-based models have been developed, including CREAMS
80 (Chemicals, Runoff and Erosion from Agricultural Management Systems; Knisel,
81 1980), AGNPS (Agricultural Nonpoint Pollution Source; Young et al., 1989), WEPP
82 (Water Erosion Prediction Project; Nearing et al., 1989), ANSWERS (Areal Nonpoint
83 Source Watershed Environment Response Simulation; Beasley et al., 1980), and HSPF
84 (Hydro-logic Simulation Program Fortran; Johanson et al., 1984). Physically-based
85 models are calibrated through empirical coefficients or exponents for practical
86 applications (Borah and Bera, 2003; Aksoy and Kavvas, 2005), and thus are highly

dependent on data accessibility (Boardman, 2006; De Vente et al., 2006), especially when applied to the assessment of large areas (Mutekanga et al., 2010). Semi-quantitative models such as PSIAC (PSIAC, 1968) and FSM (Verstraeten et al., 2003) have less strict data requirements (De Vente and Poesen, 2005; Haregeweyn et al., 2005), but their applications to large basins are still limited owing to the divergence in empirical parameters for different small basins. With the aid of 3S technology, physically-based models (Vrieling, 2006; Tian, 2010) could be used for larger areas, but new challenges arise in how to deal with the massive quantity of data. For DMMP, uncertainty resulting from the discrimination analysis needs to be further minimized.

Ni et al. (2008) proposed a Discrimination Method based on Minimum Polygons (DMMP) for assessment of soil erosion based on the overlay analysis of spatial multi-factors. An erosion index (*EI*) is used for each polygon by multiplying the normalized environmental factors by weights determined using the Analytic Hierarchy Process (Saaty, 1980). Representative polygons are selected and then clustered into reference groups according to erosion grade, whereas the others are assigned to test groups. For each reference group, a discrimination rule is derived between the soil-erosion grades of minimum polygons and their *EI*s in order to assess the soil-erosion severity level within each polygon in the test groups.

This paper proposes a smart coding system (SCS) to encode graded information on each environmental factor. Increasingly large areas are represented by multiple coded polygons derived from the overlay of coded factors,. This permits efficient assessment of the severity of soil erosion in large basins such as the Yellow River Basin.

2. Methodology

2.1. Classification and Coding Schema for Geographic Information (CCSGI)

Geographic information is often comprehensive and derived from different sources, including maps, numerical data and texts describing geographical entries. To facilitate data handling, Classification and Coding of Information (CCI) transforms geographic information into a set of coding elements via certain prescribed rules. Coding is based on information classification according to independent attributes (Figure 1). Standard methods for CCI include hierarchic classification and faceted classification (SAQSIQ, 2002). For CCSGI, it is supposed that hierarchic classification is suitable for qualitative information, whereas faceted classification is suitable for detailed quantitative information. CCSGI unites qualitative and quantitative information by applying these two classification methods together.

[Place Figure 1 here]

Hierarchic classification is widely used in many fields given that hierarchic structures are commonplace (Boulton and Wallace, 1973; Zheng, 2000; Dale and Wallace, 2005;

129 Dale et al., 2010). Figure 2 shows the dendrogram structure of a hierarchy with
130 defined levels. In hierarchic classification, the population is divided into *N* classes,
131 and then each class is further subdivided into independent refined sub-classes at the
132 next level, based on the hierarchic relationships between sub-classes and their
133 node-class. This process repeats until all terminal classes i.e. class-*k* at level-*j* (Figure
134 2) contain enumerable or numeric information that are inappropriate for hierarchic
135 classification but suitable for faceted classification. For a given level of hierarchic
136 classification, a coding template is derived that consists of the terminal classes at this
137 level. The coding template concisely conveys synthetic information concerning the
138 geographic unit, and is represented by the following set:
139
140 $$W = \{X \mid X_1, X_2, ...X_i, ..., X_T\} \quad (1)$$
141
142 where X_i is an item in the coding template and T is the dimension of the set or the
143 number of attributes considered.
144 [Place Figure 2 here]
145
146 Each item of this coding template is relatively independent and describes a single
147 attribute of the geographic unit. At different levels of the hierarchic classification, the
148 coding template changes. Therefore, this classification method adapts to different
149 scales at different levels (Dale and Wallace, 2005).

For each item quantified by enumerative or numeric information in the coding template, the faceted classification method is further used to categorize the information into a specific state or facet according to predefined partitioning rules. Each facet or state may represent several enumerable values or a range of detailed values between two thresholds. Hence, information on the population can be reduced to multi-states. Item X_i in set Ω is given as follows:

$$X_i = \{X_i \mid x_i^1, x_i^2, ..., x_i^j, ..., x_i^k\} \tag{2}$$

where x_i^j is the state j of X_i; k is the total number of states belonging to X_i.

This classification scheme is inherently able to describe the subject domain using simplified quantitative information (Prieto-Diaz, 1991; Herring, 2007). Moreover, a specific numeric code is assigned for each state/facet and considered as a substitute for the source information. As classification information, the code is much more tolerant to data deficiency and inaccuracy than the quantitative numeric information. In other words, faceted classification helps the data requirement to be fulfilled.

In short, a mass of given geographic information is partitioned into T classes by hierarchic classification rules. Subsequently, the codes are obtained by faceted

171 classification rules as follows:

172

173 $$C = \{C \mid (c_1, c_2, ..., c_i, ..., c_T), c_1 \hat{I} X_1, c_2 \hat{I} X_2, ..., c_i \hat{I} X_i, ..., c_T \hat{I} X_T\} \quad (3)$$

174

175 where c_i is the code of element X_i in set Ω.

176

177 The code value c_i is either assigned an ordered integer ranging from 1 to k, or else
178 values based on its application so as to facilitate easy expansion of the coding system
179 (and hence its usefulness). Adaptability of the code template at different levels in the
180 hierarchical classification facilitates tolerance to data deficiency and inaccuracy; in
181 other words, the CCSGI is self-adaptive at different spatial scales for data of moderate
182 scarcity in a large basin.

183

184 **2.2. Selection, Classification, and Coding of Soil Erosion Environmental Factors**

185 The CCSGI is implemented in the selection, classification and coding of soil erosion
186 environmental factors in order to complete the representation of environmental factors.
187 Although information describing the environmental factors might be scale-dependent,
188 the factors are generally classified under four main headings of climate, topography,
189 soil, and vegetation (Ni et al., 2008). Figure 3 depicts the hierarchical classification
190 scheme of environmental factors systematically selected for soil erosion. Here Level 1
191 is at the highest level, whereas Level 4 the lowest level in the hierarchy. The attributes

192 at Level 1 are more qualitative than those at lower levels. Macroscopic variables appear
193 at Level 2 corresponding to basin-scale. For example, the climate variable at Level 1 is
194 further specified as annual precipitation at Level 2 for soil loss caused by rainfall. At
195 Level 3, the topographical variables are further specified as length and gradient, and
196 slope pattern. Similarly, the vegetation could be interpreted more specifically than
197 vegetation cover at the lower levels. Attributes representing precipitation, gully density
198 and soil type may remain but be resampled at higher spatial resolution. It should be
199 noted that the rain regime is more important in small than in large basins (Nearing et
200 al 2005, Fang et al 2012).
201 [Place Figure 3 here]
202
203 Table 1 lists the faceted classification codes for each environmental factor at Level 2,
204 based on the standard released by the Ministry of Water Resource (MWR), China
205 (2008), which has been widely cited in the literature (see e.g. Shi et al., 2004; Yang et
206 al, 2005; Fu et al., 2006; Zhou et al., 2008; Liu et al., 2012). Table 1 lists the
207 multi-states and corresponding ranges of values or facets corresponding to each state.
208 For example, annual rainfall less than 300mm is coded as 2; soil erodibility of loess
209 parent material is coded as 5. This makes the categorization scheme more reliable than
210 conventional empirical methods such as simple clustering or equal division (MWR,
211 2008). Alternative methods like clustering discrimination could be used in cases where
212 standardized classifications of factors such as vegetation type, slope length and slope

213 pattern are lacking (MWR, 2008). For example, cover indices of different vegetation
214 types (SEPA, 2006) could be simply calculated and graded for further coding.
215 [Place Table 1 here]
216
217 **2.3. Comparison of Coding Sequences**
218 CCSGI produces representations of environmental factors affecting soil erosion, and
219 then SCS compares the derived codes (Figure 4). The code with information on
220 graded environmental factors in a mini-polygon indicates the severity level of soil
221 erosion in the same geographic unit.
222 [Place Figure 4 here]
223
224 For comparison, reference groups are established in terms of coding sequences of
225 environmental factors, and rapid soil-erosion assessment is undertaken as follows.
226
227 (i) Coding of Mini-polygon
228 The mini-polygon is the basic spatial geographical unit for evaluation of soil erosion
229 (Wang, 1993), and is directly derived from the overlay of environmental factors using
230 GIS (Cowen, 1988; Burrough, 1992). By coupling CCSGI with tools in ArcGIS, the
231 geographic information stored in a minimum polygon is further transformed into a
232 coding sequence that is easy to handle. Via CCSGI, geographic maps of the grades of
233 each environmental factor are generated in vector format. Using ArcGIS overlay

234 analysis, a coding-sequence map is produced that contains all graded environmental
235 factors, from which the mini-polygons are generated and coded. Detailed advice on
236 ArcGIS tools is available at ArcGIS Resource Center (http://resources.arcgis.com).
237
238 (ii) Establishment of the Reference Group
239 A sample of coded mini-polygons is used to establish the reference groups. The
240 remaining coded mini-polygons constitute the test groups. Random sampling is used
241 for large numbers of coded polygons to ensure the reference groups are representative.
242
243 (iii) Matching of Polygons in the Test Group
244 Matching of coding sequences of test and reference polygons is the key step to predict
245 the severity level of soil erosion in the mini-polygons. To measure the similarity of a
246 pair of coding sequences, a coding sequence with n bits is considered as an
247 n-dimensional vector $\mathbf{c}= (c_1, c_2, \cdots, c_j, \cdots, c_n)^T$. Then, the cosine of the vector angle
248 between two coding sequences is calculated from
249
250 $$a = \frac{\mathbf{c}_1 \mathbf{c}_2^T}{|\mathbf{c}_1||\mathbf{c}_2|} \quad (4)$$
251
252 in which $\mathbf{c}_1$, $\mathbf{c}_2$ are multi-dimensional vectors representing the two coding sequences to
253 be compared. Taking the weights of the different factors into account, equation (4)
254 becomes

$$\alpha = \frac{\sum_{i=1}^{n} w_i c_{1,i} c_{2,i}}{|\mathbf{c}_1||\mathbf{c}_2|} \quad (5)$$

in which w_i is the weight of factor X_i with respect to soil loss; and $c_{1,i}$, $c_{2,i}$ are elements of vectors $\mathbf{c}_1$ and $\mathbf{c}_2$ respectively.

A series of similarity values α (α') is acquired through comparison of the coding sequences in the test and reference groups. Consequently, similar soil erosion grades are found in the mini-polygons with maximum similarity values.

3. Assessment of soil erosion status in the Yellow River Basin

3.1. Study Areas and Data Presentation

The Yellow River Basin covers a total area of 795,000 km^2. It flows through the Loess Plateau which is experiencing severe soil erosion. As shown in Figure 5, the annual gross rate of hydraulically-induced soil erosion in 1990s exceeded 5000 t/km^2 (MWR, 2002).

[Place Figure 5 here]

Referring to CCSGI, information on environmental factors is classified into the attributes at Level 2 in Figure 3. Datasets (i) ~ (v) are described as follows:

(i) Soil-erosion information extracted from 1:1,000,000 digital map of soil-loss intensity based on the 2nd National Soil Erosion Survey conducted in the 1990s by the Ministry of Water Resources, China and used as a data source for World Soil Information (Dijkshoorn et al., 2008). Figure 5 shows the soil erosion zonation map, with 6 grades ranging from slight erosion (Grade 1) to severe erosion (Grade 6).

(ii) Daily rainfall records at 66 hydrological stations in the Yellow River Basin available from 1990 to 1999 via China Meteorological Data Sharing Service System (http://cdc.cma.gov.cn/index.jsp).

(iii) Topography data extracted from a 90m resolution DEM, provided by International Scientific & Technical Data Mirror Site, Computer Network Information Center, Chinese Academy of Sciences (http://datamirror.csdb.cn). The DEM dataset was derived from SRTM (Shuttle Radar Topography Mission) digital elevation data V4.1.

(iv) Soil data from 1:1,000,000 digital map of soil type, provided by the Institute of Soil Science in Nanjing, Chinese Academy of Sciences (http://www.soil.csdb.cn/).

(v) Vegetation data from normalized difference vegetation index (NDVI) raster maps of 8 km resolution for the period from 1990 to 1999, obtained from the Environmental and

296 Ecological Science Data Center for West China, National Natural Science Foundation
297 of China (http://westdc.westgis.ac.cn, source for this dataset is the VITO (Flemish Inst.
298 Technological Research, Belgium), http://www.vgt.vito). The data form part of the
299 GIMMS (Global Inventory Modelling and Mapping Studies)-NDVI dataset with
300 temporal scale 15-days and spatial scale 8km. The annual NDVI is the averaged value
301 within each year, from which the multiple annual NDVI is further derived.
302
303 Within the period of interest from 1990 to 1999, Dataset (i) is used for validation of
304 assessment results of SCS, whereas Datasets (ii) ~ (v) are used as input information of
305 SCS. The data are considered sufficiently accurate if they provide enough information
306 is provided for the coding of each environmental factor based on Table 1.
307
308 **3.2. Assessment Process**
309 **3.2.1. Data Processing**
310 (i) Rainfall factor: Mean annual rainfall are derived from the daily rainfall at each
311 meteorological station, and then a scatter map is created using ArcGIS with
312 corresponding information on the latitudes and longitudes of the stations. Kriging
313 interpolation is used to obtain a raster map of mean annual rainfall throughout the
314 basin.
315
316 (ii) Topographical factors: Datum values of erosion surface elevation, gully density and

317 relative height of terrain are determined using ArcGIS from the DEM (Tang and Yang,
318 2006).
319
320 (iii) Soil factor: Erodibility grades are assigned to different soil types according to the
321 classification rules listed in Table 1.
322
323 (iv) Vegetation factor: Vegetation cover (*C*) is obtained from the NDVI map by (Zhao,
324 2003)
325
326 $$C = \frac{NDVI - NDVI_{\min}}{NDVI_{\max} - NDVI_{\min}} \quad (6)$$
327
328 where $NDVI_{\min}$ and $NDVI_{\max}$ are the minimum and maximum values of *NDVI*,
329 respectively.
330
331 **3.2.2. Coding and Identification of Mini-polygons**
332 The CCSGI is used to encode the environmental factors by faceted classification.
333 Table 1 indicates how the rainfall, topography and vegetation cover factors are graded
334 according to standard classification rules. Coding maps are derived from the raw data
335 on the environmental factors. All spatial gradation data at different scales are then
336 transformed into vector format. Furthermore, all coded vector maps are overlaid and
337 the mini-polygons generated. Each mini-polygon is identified by a specific coding

sequence. The spatial accuracy of yielded polygons is determined by the minimum scale within the maps.

3.2.3. Polygon Matching

The coded minimum polygons are randomly divided into reference and test groups. For each mini-polygon within the reference group, the grade of soil erosion intensity is determined as follows. Six grades of soil-erosion intensities are classified in reference polygons according to the 1990s' survey results. Polygon matching based on coding sequences is then undertaken to determine the soil-loss intensity of the test group. Equation (4) is used to examine the similarity of the coding sequence without considering the weights of the environmental factors. Figure 6 illustrates the pre-processing, coding, and classification procedure as applied to the assessment of soil erosion in the Yellow River Basin.

[Place Figure 6 here]

3.3. Evaluation Results

The Yellow River Basin is divided into 9916 coded polygons, of which ~90% of the total area is covered by polygons each of area less than 100 km^2, and ~75% by polygons each of area less than 50 km^2. Each polygon is represented by a corresponding coding sequence generated from graded environmental factors. Figure

359 7 shows the soil erosion intensity with a sample ratio (*SR*) of 5%, i.e. ratio of the
360 number of coded polygons in reference groups to the total number of coded polygons.
361 [Place Figure 7 here]
362
363 To quantify the degree of consistency between the calculated and observed results, a
364 variable defined as area overlap ratio (*R*) is introduced as follows:
365
366 $$R_i = \frac{\sum Ac_i}{\sum A_i} \quad (7)$$

367 where R_i is the overlap ratio of the *i*-th grade soil erosion, A_i is the surveyed area of
368 mini-polygons with *i*-th grade soil erosion over the whole basin area, and Ac_i is the
369 area of mini-polygons with the same calculated and surveyed grades of soil erosion.
370
371 Figure 8(a) presents the area overlap ratios for the six soil erosion grades. The mean
372 value of *R* is about 86.1% (with a standard error of 1.2% for 8 sets of calculations) over
373 the entire Yellow River Basin, whilst the minimum value of *R* is 75% for the sixth grade.
374 The overall accuracy is enhanced by the SCS approach, as is evident by comparison
375 against the average *R* of 76% by DMMP (Ni et al., 2008) for the same basin with the
376 same input data. For the consistency ratio of each soil erosion grade in terms of the
377 number of coded polygons, the accuracy ratio is 89.1% on average. Figure 8(b)
378 depicts the detailed overlap ratios for each grade, showing that the minimum overlap
379 ratio in terms of the number of coded polygons is 76.9% for the 6^{th} grade of soil

380 erosion intensity.

381 [Place Figure 8 here]

382

383 **4. Discussion**

384 Based on a Smart Coding System, the relationship has been properly established
385 between environmental factors and soil erosion intensity. For the Yellow River Basin,
386 a sample ratio of 5% achieves an average area overlap ratio of 86.1% with standard
387 error of 1.2% over the whole study area. Moreover, the sensitivity analysis
388 demonstrates that the sample ratio/number can be reduced further, with hardly any
389 effect on prediction accuracy. Meanwhile, the modeling uncertainty also reduces
390 compared to DMMP. SCS is not only applicable to larger basins but also more
391 efficient through data compression via CCSGI.

392

393 **4.1. Sensitivity Analysis of Sample Ratio/Number**

394 A sensitivity analysis is undertaken to examine the influence of sample ratio/number on
395 the predicted results. Figure 9 shows the change of mean area overlap ratio (R) as
396 sample ratio (SR) is increased from 0.2% to 15%. At least 8 simulations are carried
397 out for each SR to avoid uncertainty from random sampling. It can be seen that R
398 increases monotonically whereas the standard error decreases with increasing SR. For
399 $SR > 5\%$, R and its standard error reach 95% and 0.5% respectively.

400 [Place Figure 9 here]

401

402 The relationship between the mean value of R and the sample number (SN) of coded
403 polygons in the reference group is investigated to test the minimum number of coded
404 polygons required for satisfactory prediction of soil loss intensity. There is a positive
405 correlation between R and SN (Figure 9). An overlap ratio of $R \sim 80\%$ is achieved for
406 $SN \sim 200$, whereas further increase of SN does not lead to any significant gain in R. To
407 reduce workload, $SN = 200$ is sufficient as a reference value.

408

409 **4.2. Uncertainty of Assessment**

410 Similarity between coded polygons is related to uncertainty in application of the SCS,
411 and is quantified using the vector cosine between each pair of coding sequences
412 derived from CCSGI. The closer to unity the cosine value, the more reliable is the
413 matching result. Figure 10 presents a histogram illustrating the percentages of coded
414 polygons with different similarity bands; the values of similarities range from 0.96 to 1
415 with the majority close to 1. This distribution of similarities implies the assessment is
416 highly reliable. SCS seems to have more advantages over discrimination analysis for
417 assessing test groups (Ni et al., 2008) through discrimination using geographical
418 information and reduction in uncertainty. A distance index, denoted $DI = \frac{|EI - EI_0|}{EI_0}$
419 where EI is the erosion index of a test polygon and EI_0 is the central value of within its
420 matched group, is now used to measure the relative distance from EI to EI_0 and hence to
421 indicate the uncertainty of the matching results. As DI approaches 0, the matching

result is more accurate (and less uncertain). Figure 11 plots the cumulative percentage of the number of *DI* values determined using discrimination analysis. Here, *DI* is generally not close to 0, with more than 50% of values greater than 0.5, and 20% greater than 1.

[Place Figure 10 here]

[Place Figure 11 here]

Information entropy is introduced to quantify the uncertainty of the assessed results derived from the DMMP and the SCS. Information entropy φ indicates the uncertainty of information *Xi* based on its probability distribution *p(Xi)* as follows (Shannon, 1948; Li and Du, 2005):

$$j = -\sum [p(X_i)\log_2 p(X_i)] \quad (8)$$

Larger information entropy means greater uncertainty. The calculated information entropies of coded-polygon *DI*s and similarities are $\varphi = 7.49$ and $\varphi = 1.33$ for DMMP and SCS respectively, confirming the higher reliability of SCS based on coding sequences.

4.3. Efficiency for Large Basins

SCS reduces data redundancy and hence promotes efficiency of data processing. For

443 example, the number of polygons in the whole Yellow River Basin is reduced by nearly
444 90% (from 81,054 in DMMP to 9916 in SCS). For a given number *N* of basin polygons
445 and a sample ratio *SR*, the number of matches has previously been calculated from
446 $N_m = SR(1- SR)N^2$. When *N* is reduced by 90%, N_m accounts for only 1.5% of the
447 original number of matches required before CCSGI is implemented. Improved
448 efficiency is to be expected as the number of polygons increases. By setting a sample
449 ratio, the reduction in the total number of polygons also leads to a decrease in the
450 number of polygons in reference group. For the Yellow River basin, only 200 coded
451 polygons in the reference group are needed as matching polygons in the test group.
452 SCS is therefore potentially useful for a cost-effective assessment of soil erosion in
453 large basins.
454
455 **5. Conclusions**
456 Efficient assessment of soil loss is essential for sustainable river basin management.
457 This paper proposes an approach based on a smart geo-coding system coupled with a
458 rapid soil loss assessment framework. The system encodes the graded environmental
459 factors in a generated polygon and thereby determines the soil erosion intensity in the
460 polygon. Following the basic assumptions underpinning SCS, the soil erosion
461 intensity values in polygons of the test group should be similar to corresponding
462 values in polygons of the reference group, provided similar coding sequences are
463 implemented. When SCS is applied to assessment of soil erosion intensity throughout

464 the entire Yellow River Basin, satisfactory agreement is reached between the expected
465 and observed results for about 86% of the total area. Sensitivity analysis indicates that
466 the number of samples in the reference groups can be greatly reduced without loss of
467 accuracy. Herein, reliable results are obtained using less than 200 reference samples
468 from the 9916 coded polygons, which implies that only 2% representative polygons are
469 required to ensure accurate assessment. SCS inherits most of the advantages of DMMP,
470 including loose data requirement. By a simple coding-sequence matching of the
471 polygons in reference and test groups, SCS significantly reduces computational load
472 and uncertainty. SCS offers an alternative method for cost-effective assessment of
473 soil loss or conservation in large river basins.
474
475 **Acknowledgments**
476 This work was supported by the National Natural Science Foundation of China with
477 Grant No.51379010. Support from Collaborative Innovation Center for Regional
478 Environmental Quality is also acknowledged
479
480 **References**
481 Aksoy, H. and Kavvas, M.L. (2005). A review of hillslope and watershed scale erosion
482 and sediment transport models. Catena. 64(2-3), 247-271.
483 doi:10.1016/j.catena.2005.08.008.
484 Ananda, J. and Herath, G. (2003). Soil erosion in developing countries: a

485 socio-economic appraisal. Journal of Environmental Management. 68(4), 343-353.
486 doi:10.1016/S0301-4797(03)00082-3.
487 Beasley, D.B., Huggins, L.F. and Monke, E.J. (1980). ANSWERS: a model for
488 watershed planning. Transactions of the ASAE. 23(4), 938-944.
489 Biesemans, J., Meirvenne, M.V. and Gabriels, D. (2000). Extending the RUSLE with
490 the Monte Carlo error propagation technique to predict long term average off-site
491 sediment accumulation. Journal of Soil and Water Conservation. 55(1), 35-42.
492 Boardman, J. (2006). Soil erosion science: reflections on the limitations of current
493 approaches. Catena. 68(2-3), 73-86. doi:10.1016/j.catena.2006.03.007.
494 Borah, D.K. and Bera, M. (2003). Watershed-scale hydrologic and nonpoint-source
495 pollution models: review of mathematical bases. Transactions of the ASAE. 46(6),
496 1553-1566.
497 Boulton, D.M. and Wallace, C.S. (1973). An information measure for hierarchic
498 classification. The Computer Journal. 16(3), 254-261.
499 Burrough, P.A. (1992). Development of intelligent geographical information systems.
500 International Journal of Geographical Information Systems. 6(1), 1-11.
501 doi:10.1080/02693799208901891.
502 Chen, T., Niu, R.Q., Li, P.X., Zhang, L.P. and Du, B. (2011). Regional soil erosion risk
503 mapping using RUSLE, GIS, and remote sensing: a case study in Miyun Watershed,
504 North China. Environmental Earth Sciences. 63(3), 533-541.
505 doi:10.1007/s12665-010-0715-z.

506 Cowen, D.J. (1988). GIS versus CAD versus DBMS: what are the differences?
507 Photogrammetric Engineering and Remote Sensing. 54, 1551-1555.
508 Dabney, S.M., Yoder, D.C., Vieira, D. and Bingner, R.L. (2011). Enhancing RUSLE to
509 include runoff-driven phenomena. Hydrological Processes. 25(9), 1373-1390.
510 doi:10.1002/hyp.7897.
511 Dale, M.B. and Wallace, C.S. (2005). Hierarchical clusters of vegetation types.
512 Community Ecology. 1(6), 57-74. doi:10.1556/ComEc.6.2005.1.7.
513 Dale, P.E.R., Dale, M.B., Dowe, D.L., Knight, J.M., Lemckert, C.J., Choy, D.C.L.,
514 Sheaves, M.J. and Sporne, I. (2010). A conceptual model for integrating physical
515 geography research and coastal wetland management, with an Australian example.
516 Progress in Physical Geography. 34(5), 605-624. doi:10.1177/0309133310369617.
517 De Vente, J. and Poesen, J. (2005). Predicting soil erosion and sediment yield at the
518 basin scale: scale issues and semi-quantitative models. Earth-Science Reviews.
519 71(1-2), 95-125. doi:10.1016/j.earscirev.2005.02.002.
520 De Vente, J., Poesen, J., Bazzoffi, P., Van Rompaey, A. and Verstraeten, G. (2006).
521 Predicting catchment sediment yield in Mediterranean environments: the
522 importance of sediment sources and connectivity in Italian drainage basins. Earth
523 Surface Processes and Landforms. 31(8), 1017-1034. doi:10.1002/esp.1305.
524 Dijkshoorn, J.A., Van Engelen, V.W.P. and Huting, J.R.M. (2008). Soil and landform
525 properties for LADA partner countries (Argentina, China, Cuba, Senegal and the
526 Gambia, South Africa and Tunisia). ISRIC and GLADA report, ISRIC-World Soil

527 Information and FAO, Wageningen.
528 Eswaran, H., Lal, R. and Reich, P.F. (2001). Land degradation: an overview. Proc. 2nd.
529 International Conference on Land Degradation and Desertification, Thailand, Khon
530 Kaen, 20-35.
531 Fang, N.F., Shi, Z.H., Li, L., Guo, Z.L., Liu, Q.J. and Ai, L. (2012). The effects of
532 rainfall regimes and land use changes on runoff and soil loss in a small
533 mountainous watershed. Catena. 99, 1-8. doi: 10.1016/j.catena.2012.07.004.
534 Fu, B.J., Zhang, Q.J., Chen, L.D., Zhao, W.W., Gulinck, H., Liu, G.B., Yang, Q.K. and
535 Zhu, Y.G. (2006). Temporal change in land use and its relationship to slope degree
536 and soil type in a small catchment on the Loess Plateau of China. Catena. 65(1),
537 41-48. doi:10.1016/j.catena.2005.07.005.
538 Haregeweyn, N., Poesenb, J., Nyssena, J., Verstraeten, G., De Vente, J., Govers, G.,
539 Deckers, S. and Moeyersons, J. (2005). Specific sediment yield in Tigray-Northern
540 Ethiopia: assessment and semi-quantitative modeling. Geomorphology. 69(1-4),
541 315-331. doi:10.1016/j.geomorph.2005.02.001.
542 Herring, S.C. (2007). A faceted classification scheme for computer-mediated
543 discourse. Language@ Internet. 4(1), 1-37.
544 Iroumé, A., Carey, P., Bronstert, A., Huber, A. and Palacios, H. (2011). GIS application
545 of USLE and MUSLE to estimate erosion and suspended sediment load in
546 experimental catchments, Valdivia, Chile. Revista Técnica de la Facultad de
547 Ingenieria Universidad del Zulia. 34(2), 119-128.

548 Jing, K., Wang, W.Z. and Deng, F.L. (2005). Soil Erosion and Environment in China.
549 Science Press, Beijing.
550 Johanson, R.C., Imhoff, J.C., Davis, H.H., Kittle, J.L. and Donigian, A.S. (1984).
551 Hydrologic Simulation Program-Fortran (HSPF): user's manual for Release 8. EPA,
552 Environmental Research Laboratory, Athens, Georgia.
553 Kinnell, P.I.A. (2010). Event soil loss, runoff and the Universal Soil Loss Equation
554 family of models: a review. Journal of Hydrology. 385(1-4), 384-397.
555 doi:10.1016/j.jhydrol.2010.01.024.
556 Knisel, W.G. (1980). CREAMS: a field scale model for Chemicals, Runoff and Erosion
557 from Agricultural Management Systems. USDA, Conservation Research Report
558 No. 26, Washington, D.C..
559 Li, H., Chen, X.L., Lim, K.J., Cai, X.B. and Sagong, M. (2010). Assessment of soil
560 erosion and sediment yield in Liao watershed, Jiangxi Province, China, Using
561 USLE, GIS, and RS. Journal of Earth Science. 21(6), 941-953.
562 doi:10.1007/s12583-010-0147-4.
563 Li, Y.D. and Du, H. (2005). Artificial intelligence with uncertainty. National Defense
564 Industry Press.
565 Liu, Y., Fu, B.J., Lu, Y.H., Wang, Z. and Gao, G.Y. (2012). Hydrological responses
566 and soil erosion potential of abandoned cropland in the Loess Plateau, China.
567 Geomorphology. 138(1), 404-414. doi:10.1016/j.geomorph.2011.10.009.
568 Meyer, L.D. (1984). Evolution of the universal soil loss equation. Journal of Soil and

569 Water Conservation. 39(2), 99-104.
570 Millward, A.A. and Mersey, J.E. (1999). Adapting the RUSLE to model soil erosion
571 potential in a mountainous tropical watershed. Catena. 38(2), 109-129.
572 doi:10.1016/S0341-8162(99)00067-3.
573 Mutekanga, F.P., Visser, S.M. and Stroosnijder, L. (2010). A tool for rapid assessment
574 of erosion risk to support decision-making and policy development at the Ngenge
575 watershed in Uganda. Geoderma. 160(2), 165-174.
576 doi:10.1016/j.geoderma.2010.09.011.
577 MWR (2002). The bulletin of soil and water loss of china. Report, Ministry of Water
578 Resources of China.
579 MWR (2008). Standards for Classification and Gradation of Soil Erosion. Ministry of
580 Water Resources of China.
581 Nearing, M.A., Foster, G.R., Lane, L.J. and Finkner, S.C. (1989). A process-based soil
582 erosion model for USDA-water erosion prediction project technology. Transactions
583 of the ASAE. 32(5), 1587-1593.
584 Nearing, M.A., Jetten, V., Baffaut, C., Cerdan, O., Couturier, A., Hernandez, M., Le
585 Bissonnais, Y., Nichols, M.H., Nunes, J.P., Renschler, C.S., Souchere, V. and van
586 Oost K. (2005). Modeling response of soil erosion and runoff to changes in
587 precipitation and cover. Catena. 61, 131-154. doi: 10.1016/j.catena.2005.03.007.
588 Ni, J.R., Li, X.X. and Borthwick, A.G.L. (2008). Soil erosion assessment based on
589 minimum polygons in the Yellow River Basin, China. Geomorphology. 93, 233-252.

590 doi:10.1016/j.geomorph.2007.02.015.
591 Oldeman, L.R. (1994). The global extent of soil degradation. In: Soil resilience and
592 sustainable land use. Wallingford, UK: CAB International, 99-118.
593 Pimentel, D., Harvey, C., Resosudarmo, P., Sinclair, K., Kurz, D., McNair, M., Crist, S.,
594 Shpritz, L., Fitton, L., Saffouri, R. and Blair, R. (1995). Environmental and
595 economic costs of soil erosion and conservation benefits. Science. 267(5201),
596 1117-1123. doi:10.1126/science.267.5201.1117.
597 Prieto-Diaz, R. (1991). Implementing faceted classification for software reuse.
598 Communications of the ACM. 34(5), 88-97.
599 PSIAC (1968). Factors affecting sediment yield and selection and evaluation of
600 measures for the reduction of erosion and sediment yield. Report of the water
601 management subcommittee, Pacific Southwest Inter-Agency Committee (PSIAC).
602 Renard, K.G., Foster, G.R., Weesies, G.A., McCool, D.K. and Yoder, D.C. (1997).
603 Predicting soil erosion by water: a guide to conservation planning with the Revised
604 Universal Soil Loss Equation (RUSLE). National Technical Information Service,
605 United States Department of Agriculture (USDA), Washington, DC.
606 Saaty, T.L. (1980). The Analytic Hierarchy Process. McGraw-Hill Company, New
607 York.
608 SAQSIQ (2002). Basic Principles and Methods for Information Classifying and
609 Coding. State Administration of Quality Supervision Inspection in Quarantine,
610 China.

611 SEPA (2006). Technical Criteria for Eco-environmental Status Evaluation. State
612 Environmental Protection Administration of China.
613 Shannon, C.E. (1948). The mathematical theory of communication. Bell System
614 Technical Journal. 27, 379-423 and 623-656.
615 Shi, Z.H., Cai, C.F., Ding, S.W., Wang, T.W. and Chow, T.L. (2004). Soil conservation
616 planning at the small watershed level using RUSLE with GIS: a case study in the
617 Three Gorge Area of China. Catena. 55(1), 33-48.
618 doi:10.1016/S0341-8162(03)00088-2.
619 Shinde, V., Sharma, A., Tiwari, K.N. and Singh, M. (2011). Quantitative determination
620 of soil erosion and prioritization of micro-watersheds using remote sensing and GIS.
621 Journal of the Indian Society of Remote Sensing. 39(2), 181-192.
622 doi:10.1007/s12524-011-0064-8.
623 Smith, D.D. and Whitt, D.M. (1948). Evaluating soil losses from field areas.
624 Agricultural Engineering. 29, 394-396.
625 Stocking, M. (1995). Soil erosion in developing countries: where geomorphology fears
626 to tread. Catena. 25(1-4), 253-267. doi:10.1016/0341-8162(95)00013-I.
627 Tang, G.A. and Yang, Q. (2006). ArcGIS geography information system space analysis
628 tutorial. Science Press, Beijing.
629 Telles, T.S., Guimarães, M.F. and Dechen, S.C.F. (2011). The costs of soil erosion.
630 Revista Brasileira de Ciênciado Solo. 35(2), 287-298.
631 doi:10.1590/S0100-06832011000200001.

632 Terranova, O., Antronico, L., Coscarelli, R. and Iaquinta, P. (2009). Soil erosion risk
633 scenarios in the Mediterranean environment using RUSLE and GIS: an application
634 model for Calabria (southern Italy). Geomorphology. 112(3-4), 228-245.
635 doi:10.1016/j.geomorph.2009.06.009.
636 Tian, H.Y. (2010). Summary of the application of "3S" techniques in monitoring soil
637 erosion. Proceedings of Symposium from Cross-Strait Environment & Resources
638 and 2nd Representative Conference of Chinese Environmental Resources &
639 Ecological Conservation Society, China, Linyi City, 91-95.
640 UNEP (2007). Global environmental outlook: environment for development (GEO-4).
641 Valletta: United Nations Environment Programme.
642 Verstraeten, G., Poesen, J., Vente, D.J., Koninckx, X. (2003). Sediment yield variability
643 in Spain: a quantitative and semi-qualitative analysis using reservoir sedimentation
644 rates. Geomorphology. 69(1-4), 315-331. doi:10.1016/S0169-555X(02)00220-9.
645 Vrieling, A., Sterk, G. and Beaulieu, N. (2002). Erosion risk mapping: a
646 methodological case study in the Colombian Eastern Plains. Journal of Soil and
647 Water Conservation. 57(3), 158-163.
648 Vrieling, A. (2006). Satellite remote sensing for water erosion assessment: a review.
649 Catena. 65(1), 2-18. doi:10.1016/j.catena.2005.10.005.
650 Wang, F. (1993). A parallel intersection algorithm for vector polygon overlay.
651 Computer Graphics and Applications, IEEE. 13(2), 74-81. doi:10.1109/38.204970.
652 Wang, W.Z. and Jiao, J.Y. (1996). Quantitative evaluation of factors influencing soil

653 erosion in China. Bulletin of Soil and Water Conservation. 16(5), 1-20.
654 Wischmeier, W.H. (1976). Use and misuse of the Universal Soil Loss Equation. Journal
655 of Soil and Water Conservation. 31(1), 5-9.
656 Wischmeier, W.H. and Smith, D.D. (1965). Predicting rainfall erosion losses from
657 cropland east of the Rocky Mountains. United States Department of Agriculture
658 (USDA), Agricultural Handbook No. 282, Washington, D.C..
659 Xu, Y.Q., Luo, D. and Peng, J. (2011). Land use change and soil erosion in the Maotiao
660 River watershed of Guizhou Province. Journal of Geographical Sciences. 21(6),
661 1138-1152. doi:10.1007/s11442-011-0906-x.
662 Yang, X., Zhang, K., Jia, B. and Ci, L. (2005). Desertification assessment in China: an
663 overview. Journal of Arid Environments. 63(2), 517-531.
664 doi:10.1016/j.jaridenv.2005.03.032.
665 Young, R.A., Onstad, C.A., Bosch, D.D. and Anderson, W.P. (1989). AGNPS: a
666 nonpoint-source pollution model for evaluating agricultural watersheds. Journal of
667 Soil Water Conservation. 44(2), 168-173.
668 Zheng, Z. (2000). Constructing X-of-N attributes for decision tree learning. Machine
669 Learning. 40(1), 35-75. doi:10.1023/A:1007626017208.
670 Zhao, Y.S. (2003). Theory and method of remote sensing application analysis. Beijing:
671 Science Press.
672 Zhou, P., Luukkanen, O., Tokola, T. and Nieminen, J. (2008). Effect of vegetation
673 cover on soil erosion in a mountainous watershed. Catena. 75(3), 319-325.

674 doi:10.1016/j.catena.2008.07.010.

675 Zingg, A.W. (1940). Degree and length of land slope as it affects soil loss in runoff.

676 Agricultural Engineering. 21, 59-64.

677

Table 1. Classification and gradation for environmental factors

Grade/ Code	Annual rainoff (mm)	Gully density (km/km^2)	Erosion base (m)	Relative Height (m)	Soil erodibility	Cover (%)
1		<1	0	<50		>90
2	<300	1~2	1000	50~200	black soils, chernozems, alpine/sub-alpine felty soils	70~90
3	300~600	2~3	4000	200~500	cinnamon soils, brown earths, yellow-brown earths	50~70
4	600~1000	3~5		500~1000	yellow earths, red earths, latosols	30~50
5	1000~1500	5~7		1000~1500	loess parent materials	10~30
6	>1500	>7		>1500	sandy soils, desert soils, loose weathering materials	<10

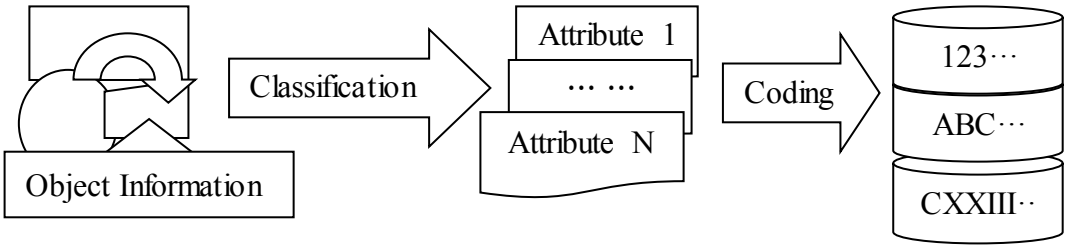


Figure 1. Classification and Coding of Information (CCI)

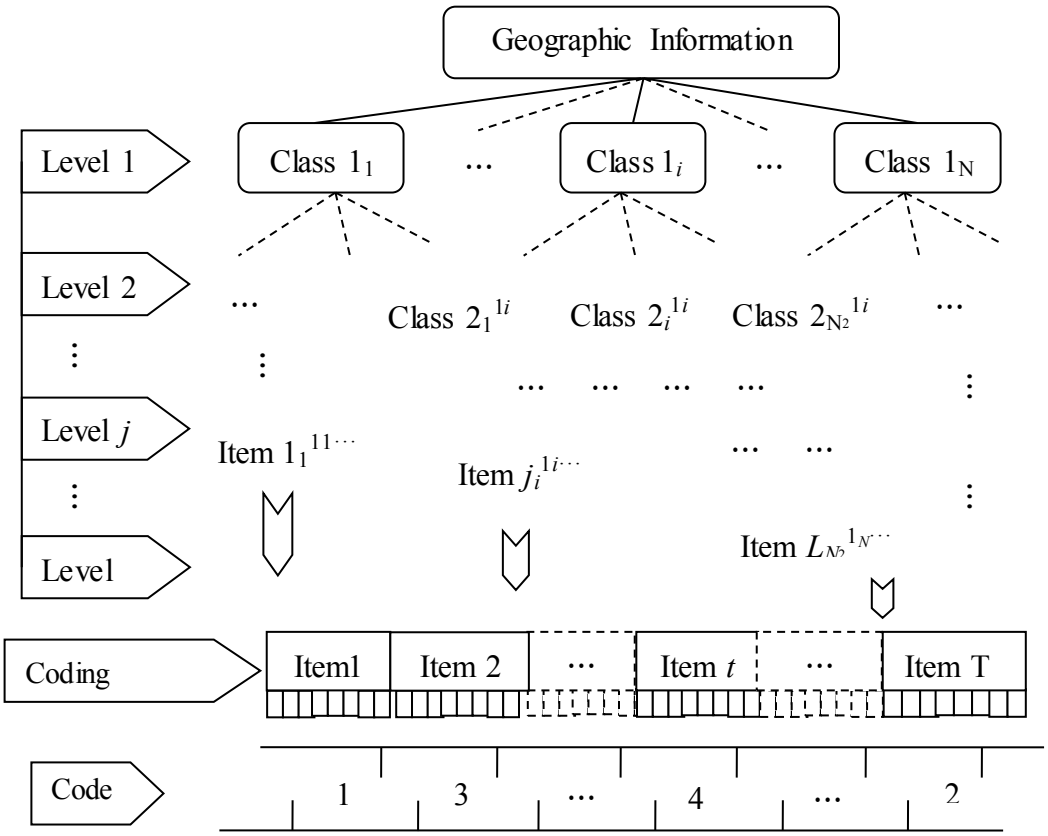


Figure 2. Classification and Coding Schema for Geographic Information (CCSGI)

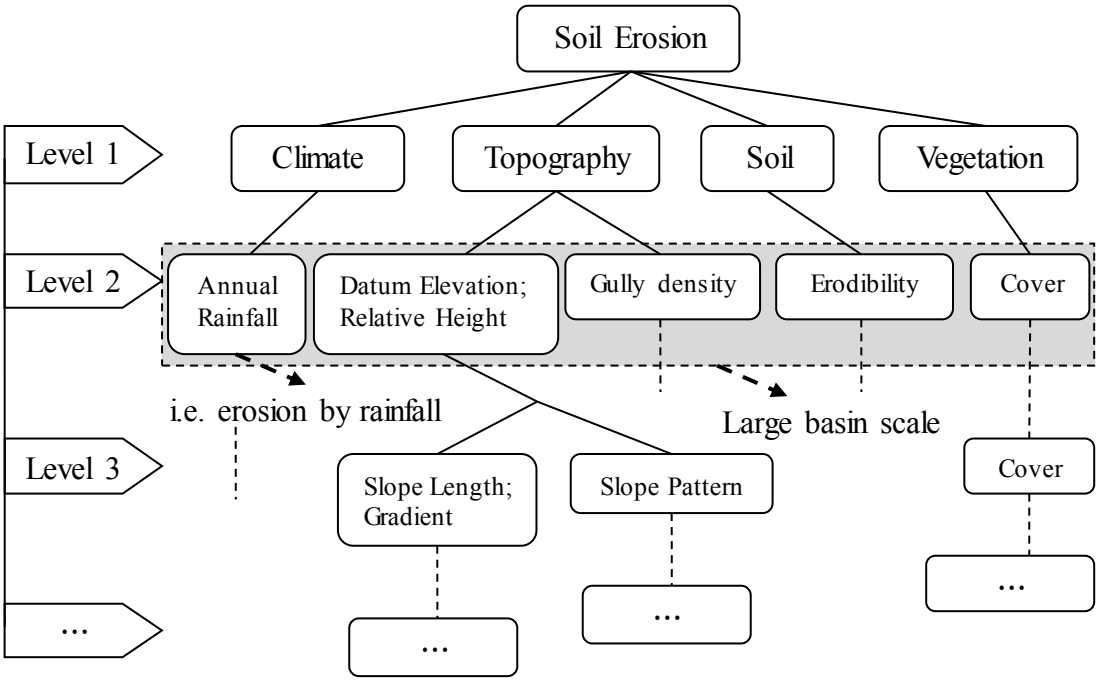
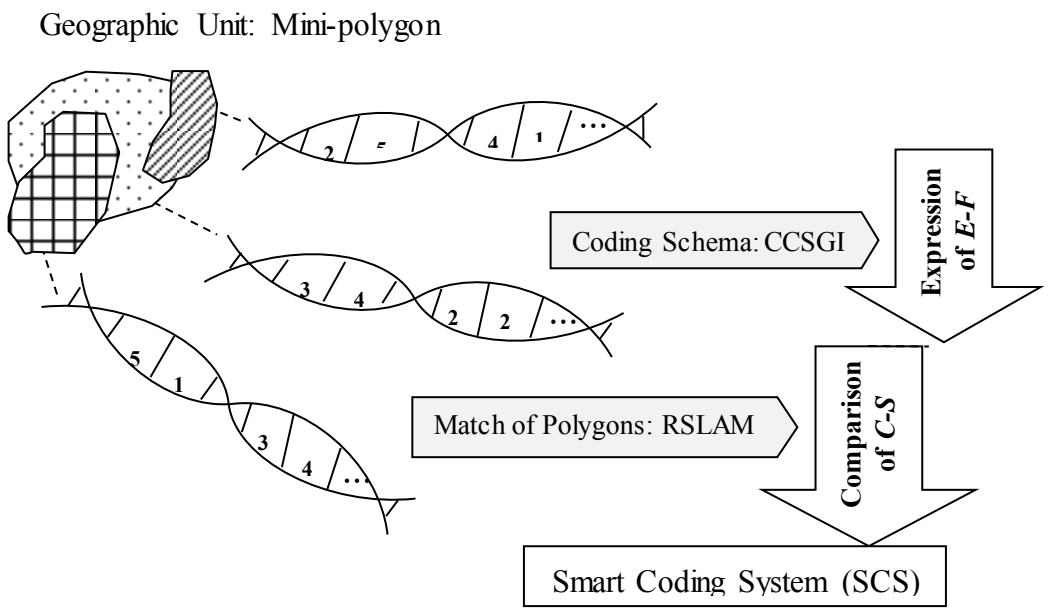


Figure 3. Hierarchical classification of factors influencing soil erosion



E-F: Environmental Factors; *C-S*: Coding Sequences.

Figure 4. From Classification and Coding Schema for Geographic Information (CCSGI) to Smart Coding System (SCS)

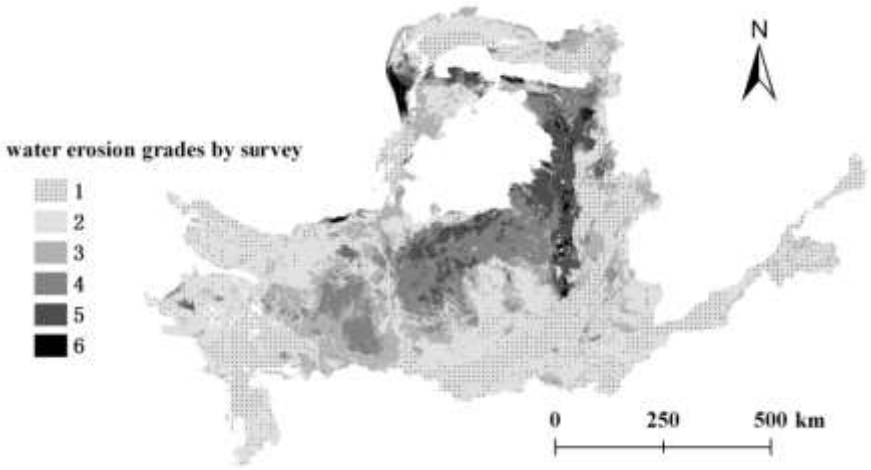


Figure 5. Water-induced soil erosion in the Yellow River Basin (MWR, 1990s)

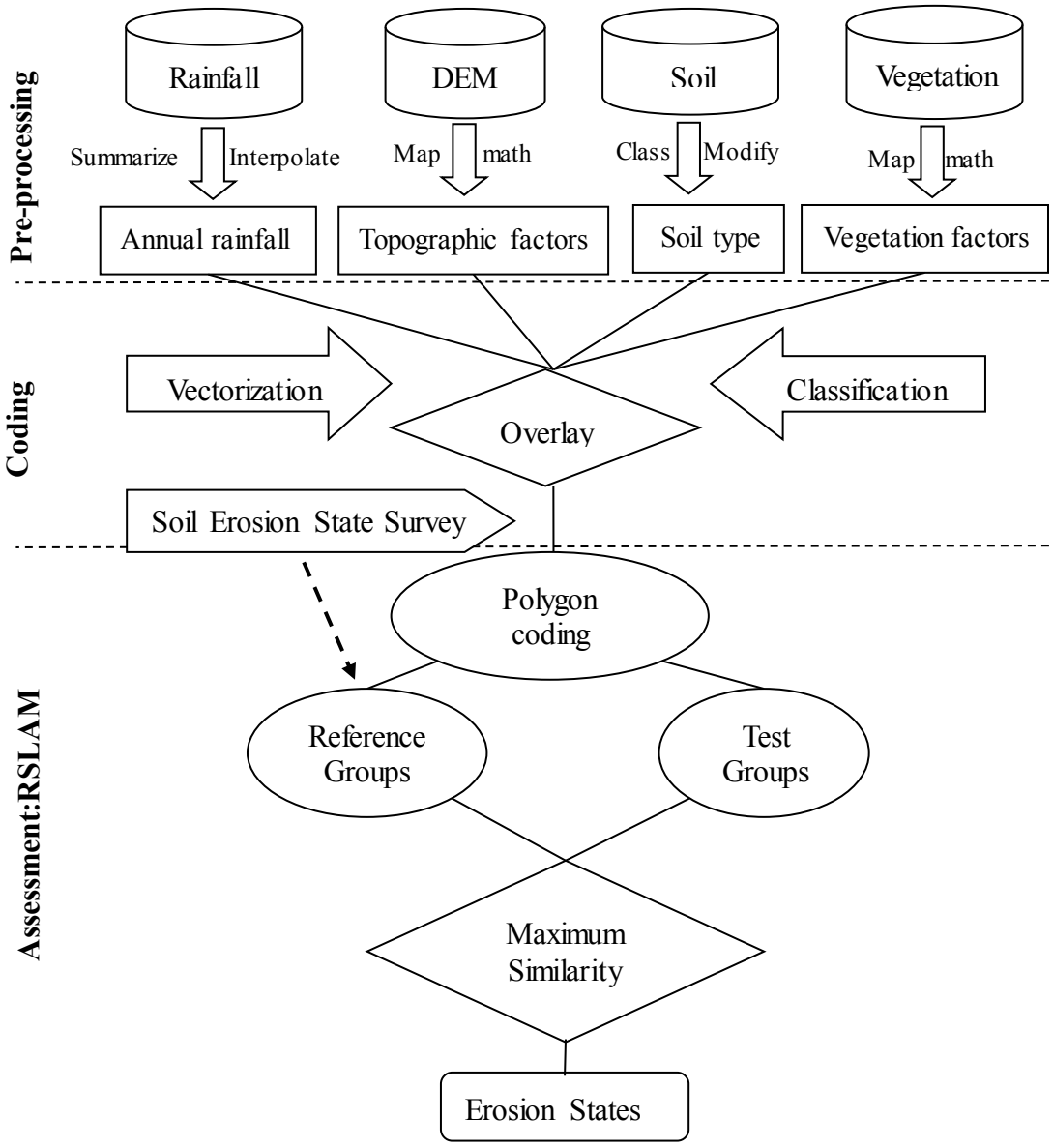


Figure 6. Smart Coding System (SCS) for soil erosion assessment in the Yellow River Basin

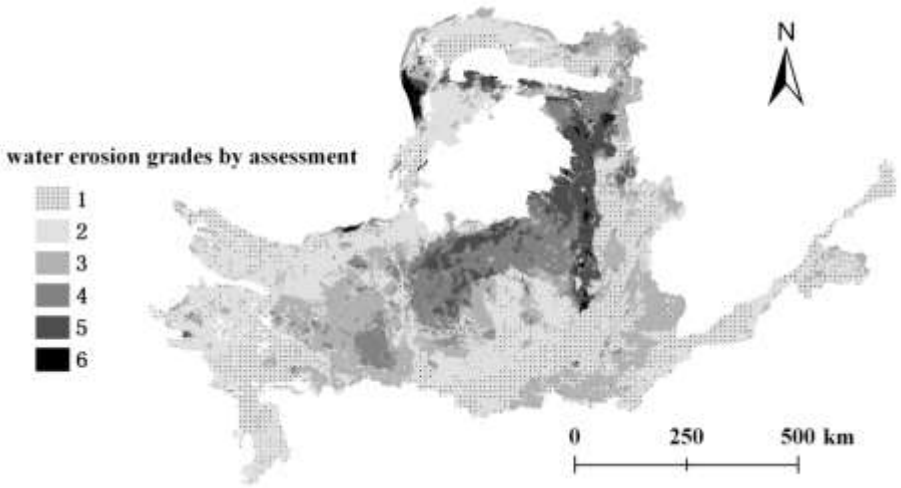


Figure 7. Spatial distribution of soil erosion intensity in Yellow River Basin from Smart Coding System (SCS)

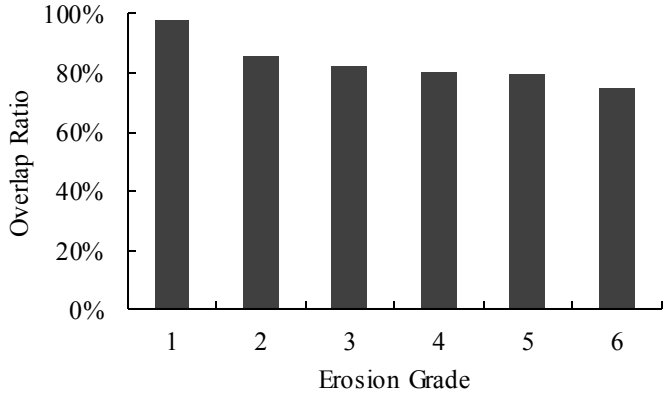


Figure 8(a) Overlap ratio of observed and calculated areas of the same soil-erosion grade

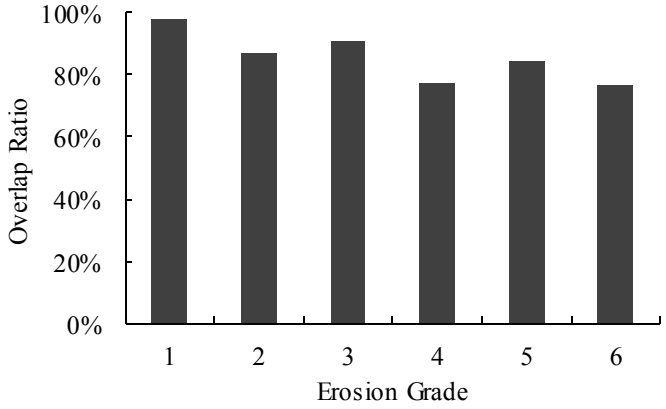


Figure 8(b) Overlap ratio of observed and calculated numbers of coded polygons of the same soil-erosion grade

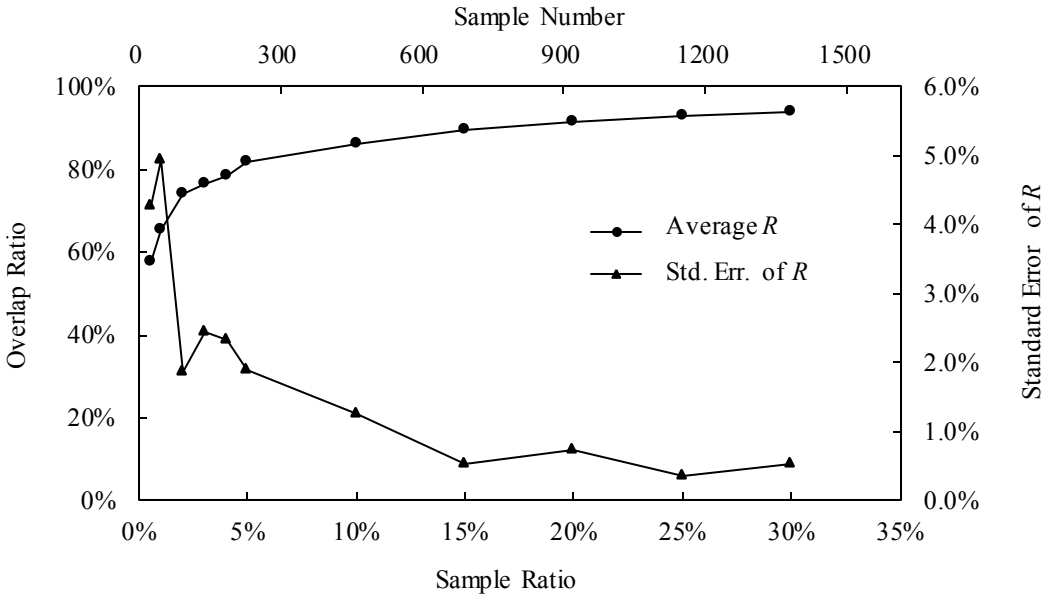


Figure 9. Sensitivity of *R* with varying sample ratio

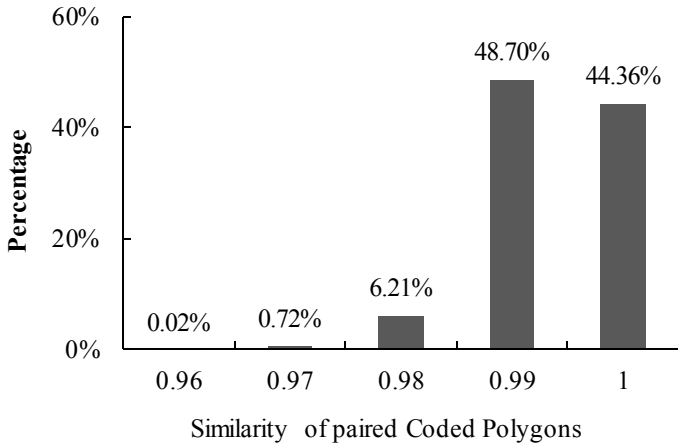


Figure 10. Percentage distribution of similarities between paired coded polygons

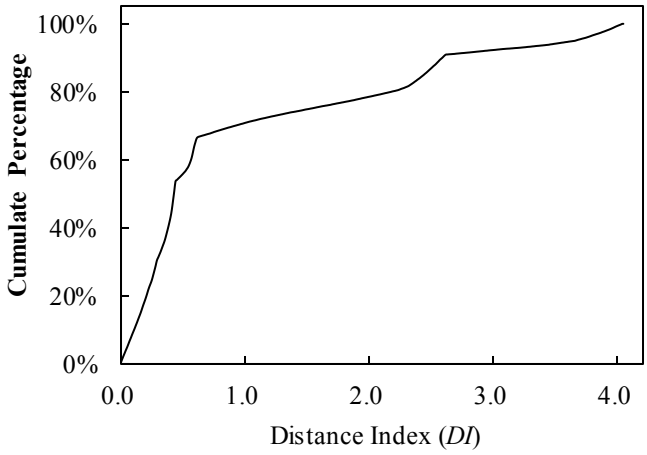


Figure 11. Percentage distribution of distance index in discrimination analysis